UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

RICHARD KADREY, et al.,

Plaintiffs,

v.

META PLATFORMS, INC.,

Defendant.

Case No. 23-cv-03417-VC (TSH)

DISCOVERY ORDER

Re: Dkt. Nos. 293, 294

The Court rules on ECF Nos. 293 and 294 as follows.

A. ECF No. 293 (Plaintiffs' Requests for Admission)

1. RFAs 3-6, 17, 20, 23, 34, 43, 45-89, 90-96

RFAs 3-6 asked Meta to admit that the datasets used to trained Llama 1, 2 and 3 and used or that will be used to train Llama 4 "included copyrighted books." For each, Meta "admits that text from a published and commercially-available version of one or more books is included in a Dataset used to train Llama [1, 2 or 3 or that may be used to train Llama 4]. As Meta lacks knowledge as to whether that text also appeared in the deposit copies submitted to the U.S. Copyright Office, which delimits what is covered by the corresponding copyright registrations for those books, Meta denies this RFA."

RFAs 17, 20 and 23 asked Meta to admit that three particular databases "contain[] copyrighted works." For each Meta responded that it "admits that text from a published and commercially-available version of one or more books appears in the [name of database]. As Meta lacks knowledge as to whether that text also appeared in the deposit copies submitted to the U.S. Copyright Office, which delimit what is covered by the corresponding copyright registrations for those works, Meta denies this RFA."

RFA 34 asked Meta to admit that the publicly available data used to train the Llama models included copyrighted works. RFA 43 asked Meta to admit that the Llama models that were trained with copyrighted material had at least in part a commercial purpose. Meta similarly admitted that the data used to train the models included text from a published and commercially-available version of one or more copyrighted works, but denied the RFAs on the ground that it lacked knowledge of whether the text appeared in the deposit copies submitted to the U.S. Copyright Office.

RFAs 45-89 asked Meta to "[a]dmit that the Books3 database contains" specific works by Plaintiffs. In each response, Meta says that it "Meta lacks information sufficient to admit or deny that [name of book], which is the subject of [name of Plaintiff's] claim and allegedly subject to copyright protection, is contained in the Books3 dataset because Meta does not possess, and Plaintiffs failed to produce, the deposit copy for [name of book] submitted to the U.S. Copyright Office, which delimits what is covered by the corresponding copyright registration. Meta admits that text from a published and commercially-available version of [name of book] is included in the dataset commonly known as Books3. As Meta lacks knowledge as to whether that text is also included in the deposit copy for this work, Meta denies this RFA."

RFAs 90-96 asked Meta to "[a]dmit that [name of book] by [plaintiff] was included in a dataset used to train Your large language models." For RFAs 90, 91, 94 and 96, Meta states that it "lacks information sufficient to admit or deny that [name of book], which is the subject of [name of plaintiff's] claim and allegedly subject to copyright protection, is contained in a dataset used to train Meta's large language models because [name of plaintiff] has not produced, and Meta does not possess, the deposit copy for [name of book] submitted to the U.S. Copyright Office, which delimits what is covered by the corresponding copyright registration. Meta admits that text from a published and commercially-available version of [name of book] is included in a dataset used to train Meta's large language models, as that term is construed above. As Meta lacks knowledge as to whether that text (or all of that text) is also included in the deposit copy for this work, Meta denies this RFA." For RFAs 92, 93 and 95, Meta does not respond at all, objecting that those books are not at issue in this action.

For all of these RFAs, Plaintiffs argue that Meta's objection that it does not know if a work is copyrighted because it does not have the deposit copy is meritless. Plaintiffs argue that publicly available information about a copyright, such as a registration or a copyright notice, is prima facie evidence of knowledge, and can be evidence of willfulness. Plaintiffs cite cases holding that a plaintiff need not produce a deposit copy to prove copyright ownership or infringement. Plaintiffs ask the Court to order Meta to drop its "deposit copy" objections. Meta, by contrast, argues that the deposit copies define the boundaries of what is copyrighted. As Meta says that it does not have the deposit copies, Meta says that it cannot determine whether the works in question are copyrighted.

Meta is wrong. All or nearly all of the books at issue that were written since the Copyright Act of 1976 took effect are copyrighted.[1] Do you know why? Because they are books. "The Copyright Act of 1976 made copyright automatic upon fixation of a work in a tangible medium, and that regime persists today, meaning mandatory deposit remains unnecessary to gain copyright." *Valancourt Books, LLC v. Garland*, 82 F.4th 1222, 1233 (D.C. Cir. 2023); *see also Georgia v. Public.Resource.Org, Inc.*, 590 U.S. 255, 275 (2020) ("Unlike other forms of intellectual property, copyright protection is both instant and automatic. It vests as soon as a work is captured in a tangible form, triggering a panoply of exclusive rights that can last over a century."); 17 U.S.C. § 102(a) ("Copyright protection subsists, in accordance with this title, in original works of authorship fixed in any tangible medium of expression, now known or later developed, from which they can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device."). It is telling that Meta's argument to the contrary relies on two cases that applied the Copyright Act of 1909. *See Skidmore v. Led Zeppelin*, 952 F.3d 1051, 1064 (9th Cir. 2020); *Structured Asset Sales, LLC v. Sheeran*, 559 F. Supp. 3d 172, 174 (S.D.N.Y. 2021).

It is true that, in general, a copyright registration is required in order to file a civil action

[1] There are some exceptions for materials that cannot be copyrighted. *See, e.g., Public.Resource.Org*, 590 U.S. at 259 (government edicts doctrine); *Feist Publications, Inc. v. Rural Telephone Service Co., Inc.*, 499 U.S. 340, 362 (1991) (telephone white pages not original or creative).

for infringement, *see* 17 U.S.C. § 411(a), and obtaining the registration triggers the deposit requirement. 17 U.S.C. § 408(a) & (b). But these RFAs do not ask about copyright *registrations*. RFAs 3-6, 17, 20, 23, 34 and 43 ask about "copyrighted books," "copyrighted works," and "copyrighted material." Half a century ago Congress amended the Copyright Act to provide that a copyright and a copyright registration are different things. RFAs 3-6, 17, 20, 23, 34 and 43 ask about copyrights, and Meta is not allowed to answer in terms of registrations. For these RFAs, Meta can deny based on its "deposit copy" objection only if it has a good faith belief that *all* of the works at issue in each RFA were written before the Copyright Act of 1976 took effect. For works written after the Copyright Act of 1976 took effect, whether the works are registered and what the deposit copies are have simply nothing to do with whether the works are copyrighted.

Further, Meta's "deposit copy" objection is frivolous as to RFAs 45-89, 90, 91, 94 and 96. Not only did those RFAs not ask about registrations, they didn't ask about copyrights either. RFAs 45-89 asked Meta to admit that the Books3 database contains particular works. And RFAs 90, 91, 94 and 96 asked Meta to admit that particular books were in the datasets used to train large language models. Meta's answers about copyright protection, registrations and deposit copies have nothing to do with the RFAs. Further, if you parse through all the legalese in Meta's responses, you can figure out that despite the uniform denials, the answer in every case is actually yes. Meta is being evasive and did not answer these RFAs as required.

The Court **GRANTS** Plaintiffs' motion to compel as to RFAs 3-6, 17, 20-, 23, 34, 43, 45-89, 90, 91, 94 and 96. The Court **DENIES** Plaintiffs' motion as to RFAs 92, 93 and 95, as Meta did not make the "deposit copy" objection in those responses, and Plaintiffs did not brief Meta's actual objection (i.e., that the particular work asked about is not at issue in this lawsuit).

2. RFAs 8, 9, 18, 21, 24, 10-13, 1

a. RFAs 8, 9, 18, 21, 24

These RFAs asked Meta to admit that it did not obtain permission from Plaintiffs or compensate Plaintiffs to do certain things, and that Meta was not authorized by all copyright owners to do certain things. In each response, Meta stated that it did not seek, obtain, *or need*

permission, to compensate, or authorization.

The Court agrees with Plaintiffs that Meta has improperly refused to answer the RFAs as written and instead rewrote them. Plaintiffs did not ask Meta for its opinion about what it needed to do. The Court **GRANTS** Plaintiffs' motion to compel as to these RFAs and **ORDERS** Meta to remove the language about what it needed or didn't need to do from its responses.

b. RFAs 10-13

These RFAs asked Meta to "[a]dmit that You have made Llama [1, 2, 3, 4] available for use by Third Parties." For RFAs 10-12, Meta states: "Meta admits that it has made Llama [1, 2, 3] available for use by Third Parties under certain circumstances and subject to certain terms and restrictions. Except as expressly admitted, Meta denies the Request." For RFA 13, Meta admits "that it currently intends to make Llama 4 available for use by Third Parties at some point in the future under certain circumstances and subject to certain terms and restrictions. Except as expressly admitted, Meta denies the Request."

Plaintiffs ask the Court to order Meta to remove the language about certain circumstances and subject to certain terms and conditions. The Court **DENIES** Plaintiffs' motion to compel as to these RFAs because these RFAs as written are vague. Meta's qualifying language is appropriate.

c. RFA 1

This RFA asked Meta to "[a]dmit that Meta created and maintains the large language models known as Llama." Meta responded that it "admits that it created a family of generative artificial intelligence ('AI') large language models known under variations of the 'Llama' name (i.e., Llama 1, Llama 2, Llama 3), which Meta released under open source licenses. Except as expressly admitted, Meta denies the Request."

Plaintiffs object to Meta's reference to releasing the LLMs under open source licenses, which they say is not responsive to the RFA. Meta argues that that the models were released under open source licenses and therefore that users can fine-tune, distill and deploy the models anywhere.

The Court **GRANTS** Plaintiffs' motion to compel as to this RFA. Meta's response answers the question about creation, but Meta's answer does not say whether or not Meta

maintains the models. Meta must answer that part of the RFA. If Meta maintains the models, but it does not exclusively maintain them, it would be an appropriate qualification to say who else also maintains them. If the open source licenses mean that others maintain the models, Meta has to say that. The current reference to open source licenses has nothing to do with the RFA.

3. RFA 7

This RFA asked Meta to "[a]dmit that You did not obtain permission or consent from the relevant copyright owners to use all copyrighted books in the Datasets used to train Llama Models." Meta responds that it "is unable to respond to this Request and on that basis denies the Request. Meta is willing to meet and confer to understand how to interpret this Request."

Plaintiffs seek an order compelling a response. Meta claims that "[i]t is wholly unclear what is meant by 'relevant copyright owners' and 'all copyrighted books.' Among other things, is this limited to registered copyrights in the US?"

The Court **GRANTS** Plaintiffs' motion to compel as to this RFA. It is perfectly clear. The relevant copyright owners are the ones who own the copyrights to the books in the datasets used to train Llama models. The copyrighted books are the ones in those datasets. Since this case arises under U.S. copyright law, the reference to copyrights means copyrights under U.S. law. And no, this RFA is not limited to registered copyrights. Half a century ago Congress rewrote the Copyright Act to extend protection to works of authorship fixed in any tangible medium of expression. Those are the copyrights this RFA is asking about.

4. Llama 5

Plaintiffs say that for the same reasons as stated in their letter brief in ECF No. 294, Meta's interpretation of "Llama Models" to exclude Llama 5 in its RFA responses is improper. Meta states that Llama 5 is far out in the future and RFAs about it are speculative and hypothetical. The Court agrees with Plaintiffs and **ORDERS** Meta to include Llama 5 in the term "Llama Models." Of course, if the answer to any particular RFA is different for Llama 5 than it is for the earlier Llama models, Meta can say that in its responses.

B. ECF No. 294 (Plaintiffs' Requests for Production)

1. Definitions

a. "Llama Models"

Plaintiffs argue that Llama 5 is relevant. They do not make that argument in connection with any particular RFP. Meta's position is unclear. Meta says that Llama 5 is even more nascent than Llama 4, which is still under development. But Meta also says that it is not withholding responsive materials on the ground that they concern Llama 5. Meta says that its search terms to identify responsive documents are not limited to specific Llama models.

The Court agrees with Plaintiffs that Llama 5 is relevant and **ORDERS** that Meta may not decline to produce otherwise responsive documents on the ground that they concern Llama 5.

b. "Shadow Datasets"

RFP 81 asked for "[a]ll Documents and Communications related to the decision to use Shadow Datasets for training Llama Models." Plaintiffs define the term "Shadow Datasets" as the type of databases described in paragraph 37 of the Complaint, which alleges "Bibliotik is one of a number of notorious 'shadow library' websites that also includes Library Genesis (aka LibGen), Z-Library (aka B-ok), Sci-Hub, and Anna's Archive. The books and other materials aggregated by these websites have also been available in bulk via torrent systems. These shadow libraries have long been of interest to the AI-training community because of the large quantity of copyrighted material they host. For that reason, these shadow libraries are also flagrantly illegal."

Plaintiffs complain that Meta's response to RFP 81 is limited to the specific examples of Shadow Datasets that Plaintiffs identified, when they said "including but not limited to" the specific examples. Meta says it doesn't know what the "Shadow Datasets" are other than as identified by Plaintiffs. Meta adds that it has disclosed all datasets that were used for Llama 1, 2 and 3 and has produced documents sufficient to show various other datasets that have been or are being considered for use with future models.

Rule 34 requires an RFP to "describe with particularity each item or category of items to be inspected . . ." Fed. R. Civ. Proc. 34(b)(1)(A). Whether a dataset is a Shadow Dataset appears to depend on whether it is a notorious shadow library, which sounds like a matter of opinion. And

the question is not whether Meta thinks that the dataset is notorious, because Plaintiffs seem to be asking Meta about datasets that other people think are notorious. Nor is this a situation where a litigant is being asked to use common sense and interpret terms the way a regular person would, as the typical person has likely never heard of these things and probably has no opinions about them. There is apparently some specialized community in which people have opinions about which datasets are notorious shadow library websites. The Court thinks that the "including but not limited to" portion of the definition of Shadow Datasets does not describe anything with particularity. It's too much to ask Meta to guess what opinions some specialized community has about certain websites. As Meta has disclosed all the datasets that were used for Llama 1, 2 and 3, Plaintiffs can tell Meta which of those are Shadow Datasets. Plaintiffs' motion to compel is **DENIED** on this issue.

c. "Agreements"

RFP 91 asked for "[a]ll Documents and Communications related to any licensing agreements for Llama Models, including terms, conditions, and consideration." RFP 130 asked for "[a]ll Documents and Communications, including discussions, deliberations, or negotiations related to any actual, proposed, or contemplated licensing agreements (even if never written or executed) for AI training data, including any actual, proposed, or contemplated terms, conditions, and consideration."

Plaintiffs complain that Meta interprets the word "agreement" in both RFPs to mean only written contracts. Meta states that from the outset of discovery, Plaintiffs have defined "agreements" broadly to encompass, among other things, any oral contract, arrangement or understanding, whether formal or informal. Meta has objected and complains that Plaintiffs have not explained how there will be documents about oral agreements or understandings, let alone informal ones.

The Court **GRANTS** Plaintiffs' motion to compel on this issue. Meta cannot unilaterally limit the meaning of "agreements" to written contracts. It can also include oral contracts, arrangements or understandings, whether formal or informal. If there are any such things, internal company communications could easily refer to them. Plaintiffs do not need to prove the existence

of something to ask for it in discovery.

2. Relevance Objections

a. RFP 114

This RFP seeks documents related to Meta's partnerships with celebrities whose voices are used to train Meta's AI chatbot. Plaintiffs say these documents are relevant to fair use. Meta says these documents are irrelevant because they relate to publicity rights. The Court agrees with Meta. The use of celebrity likenesses, including their voice, is governed by the right of publicity; the subject matter of these documents is distinct from copyright. The Court **DENIES** Plaintiffs' motion to compel as to RFP 114.

b. RFP 115

This RFP seeks documents related to Meta's decision not to release its newest AI models to the European Union, which Plaintiffs say is because of EU data privacy regulations. The Court **DENIES** Plaintiffs' motion to compel as to RFP 115. Meta's response to EU regulations is a large subject that will sweep in many documents, but it has little or nothing to do with this case, so this discovery is not proportional to the needs of the case.

c. RFP 117

This RFP seeks documents about Meta's decision not to publicly disclose the databases it uses to train more recent Llama models. Plaintiffs says this decision was likely made because the training databases contain pirated versions of copyrighted works. Meta says this decision has no relevance to the case, and that Plaintiffs' theory for why this decision was made is conclusory speculation.

The Court **DENIES** Plaintiffs' motion to compel RFP 117. Meta's decision not to publicly disclose the databases it uses to train Llama models is not by itself relevant to the case. The suggestion that responsive documents will yield damaging admissions that would be relevant to the case is entirely speculative.

3. "Sufficient to Show" Limitations

a. RFP 91

This RFP sought "[a]ll Documents and Communications related to any licensing

agreements for Llama Models, including terms, conditions, and consideration." Meta's response is limited to documents "sufficient to show" any agreements entered by Meta to license use of the Llama models. Plaintiffs object to the "sufficient to show" limitation.

The Court agrees with Meta that this RFP is about licensing agreements for Llama models, such as licenses under which the models were distributed, or any other licenses that confer rights or entitlements to anyone outside of Meta concerning Llama models. However, the Court thinks that the full scope of this RFP 91 is relevant and proportional to the needs of the case. Discussions of agreements and the negotiation of agreements concerning the licensing of Llama models is relevant, not just the final agreements. These communications may shed light on how or why the agreements were developed, which may not be evident from the agreements themselves. This information all relates to how Meta may have benefited or attempted to benefit from the allegedly infringing conduct. Accordingly, the Court **GRANTS** Plaintiffs' motion to compel as to RFP 91.

b. RFPs 94, 100, 101

These RFPs sought "all documents and communications" concerning certain financial information about the Llama models. The Court agrees with Meta that these RFPs are overbroad and that documents "sufficient to show" the requested financial information are sufficient. Accordingly, the Court **DENIES** Plaintiffs' motion to compel as to these RFPs.

c. RFP 106

This RFP asked for "[a]ll Documents and Communications related to Meta guidelines for including or excluding copyrighted material from data used to train Llama Models." The Court agrees with Plaintiffs that this RFP seeks important information that is highly relevant to the case, and therefore Meta's "sufficient to show" limitation is improper. The Court is also concerned that if there is a dispute about whether Meta complies with its own guidelines, a "sufficient to show" response lets Meta pick and choose which documents to produce and may let Meta choose documents that support its "official" policy and omit departures from the guidelines. Further, because any searches are limited to identified custodians and non-custodial sources, despite the wording of the RFP, it's not true that Meta actually has to produce "all" responsive documents – just the responsive documents within the scope of what Meta is searching. Accordingly, the Court

GRANTS Plaintiffs' motion to compel as to RFP 106.

d. RFP 123

This RFP asked for "[a]ll non-privileged Communications with third parties regarding Plaintiffs' allegations and Meta's actual or contemplated defenses, including fair use." Meta stated that it would produce documents "sufficient to show" those communications.

This RFP seems to ask for relevant documents, and the Court does not understand what the "sufficient to show" limitations means. Meta also does not discuss this RFP in its section of the letter brief, waiving the issue. The Court **GRANTS** Plaintiffs' motion to compel as to RFP 123.

4. Commits and Pull Requests

RFP 120 requested "[a]ll Documents and Communications, including source code, relating to actual or contemplated source code changes within Llama Models, including source code 'commits' and 'pull requests.'" Plaintiffs move to compel the production of the commits and pull requests as text files. Meta says it will do so. Accordingly, the Court **ORDERS** Meta to produce the commits and pull requests as text files.

5. Privilege

RFP 125 asked for "[d]ocuments sufficient to show the identity of persons included on the following 'list serv' accounts identified on Meta's privilege log(s): 'GenAI Trust & RAI,' 'GenAI LLM Research,' 'GenAI - Leadership,' 'MPI in house.'" These are apparently listservs referenced in Meta's privilege logs. RFP 126 asked for "[a]ll Documents and Communications related to Meta's decision to use the SRT for obtaining or relaying legal advice." SRT is apparently a tool used by Meta employees for communicating with counsel.

Meta argues that RFP 125 is discovery into Meta's privilege log, and therefore concerned "existing written discovery" (ECF No. 279 at 2) and was subject to an October 23, 2024 deadline to raise with the Court. ECF No. 238. The Court disagrees. While it is true that RFP 125 concerns Meta's privilege log, Plaintiffs are moving on Meta's November 8, 2024 RFP responses, which did not exist at the time the order at ECF No. 238 was issued. Meta clearly thinks that disputes as to the other RFPs in Plaintiffs' fifth set of RFPs are timely under the scheduling order; there is no reason why RFP 125 would be different. On the merits, the Court declines to order this

information produced. This is discovery about discovery, and there is no good reason for it. The identity of the people included on the listservs would show who might or could have provided legal advice, but Plaintiffs do not explain how that would enable them to assess any particular claim of privilege.

RFP 126 seeks irrelevant information. It doesn't matter why Meta decided to use the SRT for obtaining or relaying legal advice.

Plaintiffs' motion to compel is **DENIED** as to RFPs 125 and 126.

* * *

In summary, Plaintiffs' motions to compel at ECF Nos. 293 and 294 are **GRANTED IN PART** and **DENIED IN PART** as stated above.

IT IS SO ORDERED.

Dated: December 6, 2024

THOMAS S. HIXSON
United States Magistrate Judge

United States District Court
Northern District of California